SR-CL-DMC: P-frame coding with Super-Resolution, Color Learning, and Deep Motion Compensation

Man M. Ho
Hosei University
Tokyo, Japan
man.hominh.6m@stu.hosei.ac.jp

Jinjia Zhou
Hosei University and JST, PRESTO
Tokyo, Japan
jinjia.zhou.35@hosei.ac.jp

Gang He
Xi'dian University
Xi'an, China
ghe@xidian.edu.cn

Muchen Li
Hosei University
Tokyo, Japan
muchen.li.42@hosei.ac.jp

Lei Li
Xi'dian University
Xi'an, China
l849747827@foxmail.com

Abstract

This paper proposes a deep learning based video coding framework to greatly increase the compression ratio and keep the video quality by efficiently leveraging the information from a reference. In the encoder, the input frame is compressed by down-sampling to a lower resolution, eliminating color information, and then encoding residual between the current frame and the reference frame using Versatile Video Coding (VVC). The decoder consists of two main parts: Super-Resolution with Color Learning (SR-CL), and Deep Motion Compensation (DMC). For the SR-CL part, we adopt Restoration-Reconstruction Deep Neural Network to firstly restore the missing information from compression at low resolution and compression without color. And then, the sampling degradation at high-resolution is compensated. For the DMC part, we adopt recursive-feedback architectures to propose an optical flow estimation and refinement using Dilated Inception Blocks. As a result, the work achieves 64*:1 compression ratio with* $41.81/41.34$ *dB PSNR and* $0.9959/0.9962$ *MS-SSIM on the validation/test set provided by the CLIC P-frame track challenge.*

1. Introduction

Recently, deep learning solved many problems in computer vision such as colorization [16, 12, 8, 18, 14], super-resolution [3, 4, 10, 15, 19, 20, 2, 13], frame interpolation [9, 1], and so on. It brings promising approaches for video compression. Different from compressing an image, the temporal information helps enhance performance via leveraging neighboring frames. In this paper, we solve a problem in P-frame coding, which exploits the useful information from the reference frame. In the encoder, we remove the color information and down-sample the input frame. Furthermore, Versatile Video Coding (VVC) codec with the Low Delay P configuration is applied to only compress the residual between the down-sampled gray-scale input frame and the reference frame. In the decoder, we design several networks to 1) learn and restore the color information; 2) increase the resolution; 3) compensate for the compression degradation ((e.g., ringing artifacts, distortion) and re-sampling degradation (e.g., bicubic degradation).

2. The Proposed Video Coding Framework

2.1. Overall Concept

As shown in Figure 1, in the encoder, given two high-resolution frames including a reference frame (frame 1) and a frame be compressed (frame 2), $HR_1, HR_2 \in R^{H\times W\times 3}$. Note that each frame has one luminance (Y) component and two chrominance (U and V) components. We firstly eliminate the color information by removing 2 chrominance components to get gray-scale high-resolution frames $HR_1, HR_2 \in R^{H\times W\times 1}$. And then, we further apply $2\times$ down-sampling to get the low-resolution gray-scale frames $LR_1, LR_2 \in R^{H//2\times W//2\times 1}$. Finally, we only encode the residual between the two frames using VVC with Low Delay P configuration. In the decoder, we firstly get a decoded low-resolution gray-scale current frame $DLR_2 \in R^{H/2\times W/2\times 1}$. After that, our deep neural network continuously leverages the colorful $LR_1 \in R^{H//2\times W//2\times 3}$, which is down-sampled reference frame, to restore the missing information from compression and color elimination, and compensate the quality via temporal information, and reconstruct the image as the target high-resolution $\hat{HR}_2 \in$

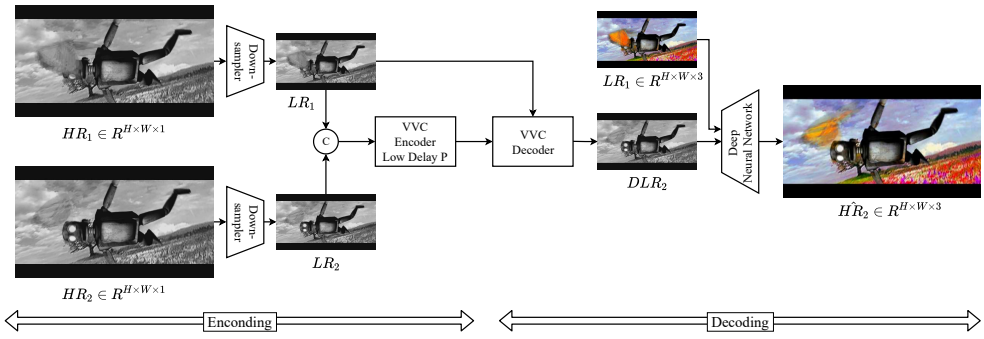

Figure 1. Overall concept.

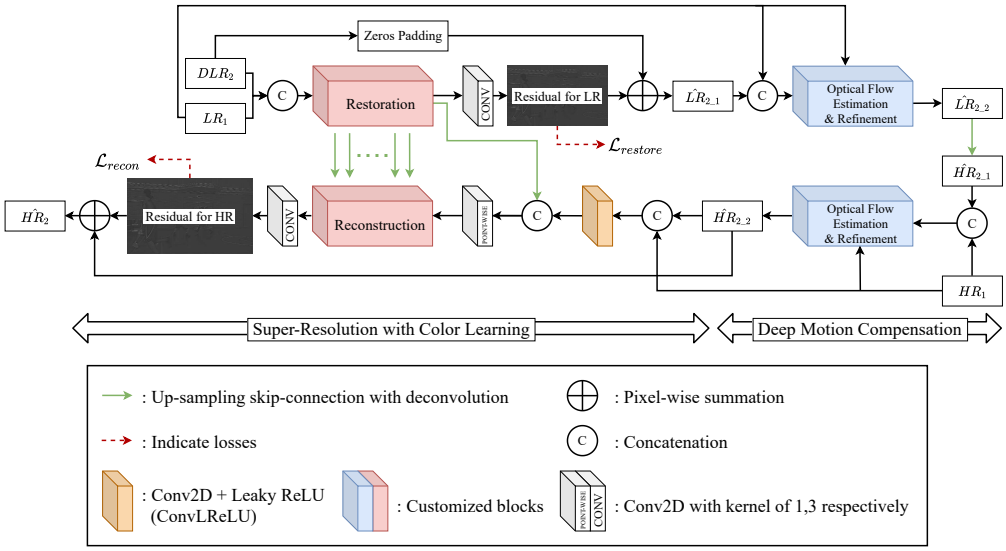

Figure 2. The proposed network SR-CL-DMC

$R^{H\times W\times 3}$. This design is applied for efficient inter coding. If the scene changes, intra coding or another inter-intra mixed coding will be applied. Therefore, for the test set, we add a Multi-Scale Structural Similarity (MS-SSIM) based selector to decide the proposed inter coding system or not.

The proposed deep neural network has two main components: Super-Resolution with Color Learning (SR-CL) and Deep Motion Compensation (DMC). As shown in Figure 2the workflow is: 1) A restoration network in SR-CL removes compression degradation and colorize DLR_2 to have $\hat{LR}_{2\_1} \in R^{H\times W\times 3}$ by learning the color and residual using the colorful LR_1. 2) DMC leverages the concatenation of $\hat{LR}_{2\_1}$ and the colorful LR_1 to estimate the optical flow, compensate missing information, and synthesize $\hat{LR}_{2\_2}$. 3) The refined $\hat{LR}_{2\_2}$ is up-sampled by a deconvolution to have $\hat{HR}_{2\_1}$. The $\hat{HR}_{2\_1}$ is compensated by DMC again to synthesize $\hat{HR}_{2\_2}$ using the colorful HR_1. 4) Finally, a reconstruction network in SR-CL reconstructs $\hat{HR}_{2\_2}$ to generate the target frame $\hat{HR}_2$.

The advantages of this work are summarized as follows: 1) The compression ratio is greatly improved by down-sampling and color elimination. The quality is kept by applying learning based super-resolution, colorization, and degradation removing. 2) The SR-CL can not only compensate for the missing information from compression but also restore the color information by using the reference frame. 3) In the DMC, the proposed flow network makes the diversity of the receptive field to enhance flow estimation. Also, the refinement network consolidates the output of the flow network using a feed-back network with several recursive blocks. They are designed for saving weights, so it does not cost a large memory while running.

2.2. Super-Resolution with Color Learning (SR-CL)

This work applies networks to learn the color and enhance the quality of the current frame by leveraging the reference frame. Based on the restoration-reconstruction ar-

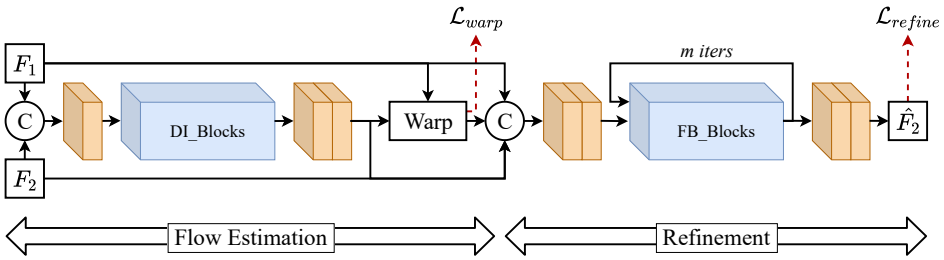

Figure 3. Our Optical Flow Estimation and Refinement.

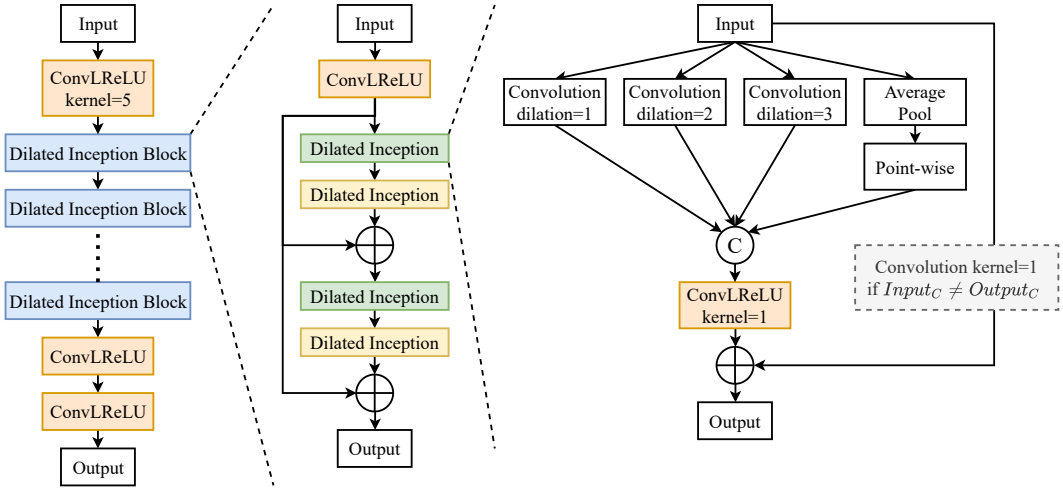

Figure 4. Dilated Inception Blocks for Optical Flow Estimation.

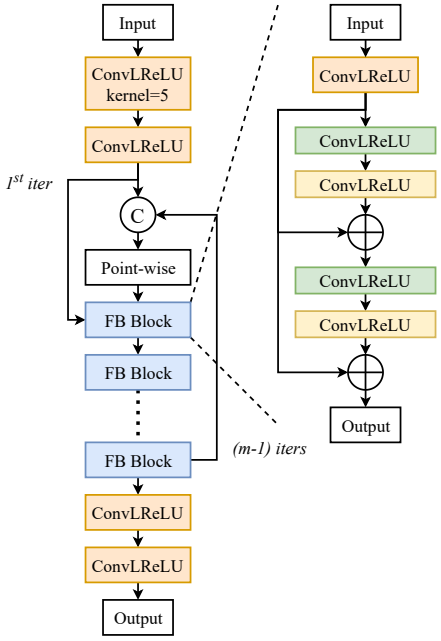

Figure 5. Feed-back (Iterative) Blocks for Refinement.

chitecture [5, 6], we design a SR-CL component to solve two problems: Super-Resolution and Colorization. SR-CL contains two networks: a restoration network learns the color and residual at low-resolution at the beginning. And then, a reconstruction network learns the final residual at high-resolution. Following its architecture, we optimize two losses $\mathcal{L}_{restore}$, $\mathcal{L}_{recon}$, as shown in Figure 2. The length of our restoration and reconstruction network is 10.

2.3. Deep Motion Compensation (DMC)

After restoration and before reconstruction, the DMC aims to enhance the quality of the restored target frame by leveraging the reference frame. We design specific networks for Flow Estimation and Refinement, respectively. Firstly, the Flow Estimation network receives two consecutive frames $F_1 \in \{LR_1, HR_1\}$, and $F_2 \in \{\hat{LR}_{2\_1}, \hat{HR}_{2\_1}\}$, as concatenation to synthesize the optical flow $OF^*_{2\to1}$ with * indicating for low-resolution or high-resolution, then warp F_1 with $OF_{2\to1}$ to generate $\hat{F}_2 \in \{\hat{LR}_{2\_2}, \hat{HR}_{2\_2}\}$, which is expected to be the original version of F_2 such as LR_2, HR_2. Afterwards, the Refinement network leverages $[F_1, F_2, \hat{F}_2, OF^*_{2\to1}$ to refine and synthesize the final $\hat{F}_2$. In DMC, we optimize the $\mathcal{L}_{warp}$ for warping optical flow, $\mathcal{L}_{refine}$ for the refinement, as shown

in Figure 3.

Flow Estimation. consists of a ConvLReLU 5×5 (Convolution followed by an activation function Leaky ReLU), five Dilated Inception Blocks (DI Blocks), and two ConvLReLU 3×3 in inference order. DI Block contains a ConvLReLU 3×3, a recursive block running two iterations. In the recursive block, we design a simple dilated inception which contains three convolutions with dilation=$\{1, 2, 3\}$ and an average pooling. Each dilated convolution/pooling in dilated inception observes a specific receptive field and will be selected by a point-wise ConvLReLU 1×1, as shown in Figure 4.

Refinement. consists of ConvLReLU 5×5, ConvLReLU 3×3, five Feed-Back Blocks (FB Blocks) looping m iterations, and two ConvLReLU 3×3. FB Block contains a ConvLReLU 3×3 and a recursive block including two ConvLReLU 3×3 running two iterations, as shown in Figure 5.

2.4. Loss Function

We adopt [6] and use the restoration loss $\mathcal{L}_{restore}$ for low-resolution and reconstruction loss for high-resolution $\mathcal{L}_{recon}$ as shown in Figure 2. In predicting optical flow and warping to generate the target frame, the warped frames are observed under $\mathcal{L}_{warp}$, and the loss $\mathcal{L}_{refine}$ is for the refined frames at both high-resolution and low-resolution stages, as shown in Figure 2, 3. Also, we add the loss function $\mathcal{L}_{ms\text{-}ssim}$ to optimize the Multi-Scale Structural Similarity (MS-SSIM) error between HR_2 and $\hat{HR}_2$. The details of our loss function are described in our supplemental document.

2.5. Data Preparation, Augmentation, and Training Details

The network is trained on the User Generated Content (UGC) dataset. To select the suitable pairs for training from 447,290 frames in the dataset, we firstly use the LiteFlowNet [7] to predict the optical flow of the possible pairs and eliminate the pairs have lower 5% pixels moving 3 pixels. Afterward, we use the perceptual metrics from [17], which efficiently estimate the scene changes and very large motion between two frames, to finally extract 80,000 pairs having perceptual distance lower than 0.6 for training. Please check our supplemental documents for further information about the perceptual distance on the UGC dataset. All training images are resized to 1280×720 as high-resolution, 640×360 as low-resolution using bicubic interpolation. To make various training samples, we use a random crop with a size of 256×256, random flip with both horizontal and vertical way, and random rotation with degrees $\in \{0, 90, 180, 270\}$. In details, we train the models with Adam optimizer [11] with learning rate of 0.0001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, the batch size of 2 on Tesla V100.

Figure 6. Illustrations of our results. The MS-SSIM score of Y, U, V components respectively are shown bellow each result. See more in our supplemental document.

3. Experimental Results

We, as the team "Man", participate in the CLIC2020 challenge on the P-frame track compressing a set of frames under the target 0.075 bpp including model size. Our method leverages the uncompressed previous frame as a reference to compensate for the missing information through down-sampling, color elimination, and video compression. As a result, as shown in Figure 6, our method performs well on the first sample. Our work can synthesize the target frame with high quality and correct colors, even the occlusions happen in the *row 1, 2, 3*; especially the *"solider"* in *row 3* who doesn't show up in the first frame. Furthermore, our method can handle the blurry inputs well as the sample in *row 4*. Please check our supplemental documents for more interesting results with high resolution. Objectively, we achieve PSNR/MS-SSIM as $41.811/0.9959$ on validation set, $41.34/0.9962$ on test set with decoder size of $4,685,293$ bytes. As mentioned about weights saving, our decoder size is smallest and smaller $22.5\times$ than the size of the largest decoder in the top 10 with competitive performance.

4. Acknowledgements

This work is supported by JST, PRESTO Grant Number JPMJPR1757 Japan.

References

[1] Wenbo Bao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Depth-aware video

frame interpolation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3703–3712, 2019. 1

[2] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11065–11074, 2019. 1

[3] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015. 1

[4] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *European conference on computer vision*, pages 391–407. Springer, 2016. 1

[5] Minh-Man Ho, Gang He, Zheng Wang, and Jinjia Zhou. Down-sampling based video coding with degradation-aware restoration-reconstruction deep neural network. In *International Conference on Multimedia Modeling*, pages 99–110. Springer, 2020. 3

[6] Man M. Ho, Jinjia Zhou, and Gang He. Rr-dncnn v2.0: Enhanced restoration-reconstruction deep neural network for down-sampling based video coding. *arXiv preprint arXiv:2002.10739*, 2020. 3, 4

[7] Tak-Wai Hui, Xiaoou Tang, and Chen Change Loy. Liteflownet: A lightweight convolutional neural network for optical flow estimation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8981–8989, June 2018. 4

[8] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Let there be color! joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Transactions on Graphics (ToG)*, 35(4):1–11, 2016. 1

[9] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9000–9008, 2018. 1

[10] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016. 1

[11] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[12] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *European Conference on Computer Vision*, pages 577–593. Springer, 2016. 1

[13] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3147–3155, 2017. 1

[14] Bo Zhang, Mingming He, Jing Liao, Pedro V. Sander, Lu Yuan, Amine Bermak, and Dong Chen. Deep exemplar-based video colorization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1

[15] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Learning a single convolutional super-resolution network for multiple degradations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3262–3271, 2018. 1

[16] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016. 1

[17] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 4

[18] Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S Lin, Tianhe Yu, and Alexei A Efros. Real-time user-guided image colorization with learned deep priors. *ACM Transactions on Graphics (TOG)*, 9(4), 2017. 1

[19] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 286–301, 2018. 1

[20] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018. 1